

# Machine Learning Techniques for Stackelberg Security Games: a Survey

Giuseppe De Nittis, Francesco Trovò

Politecnico di Milano

Dipartimento di Elettronica, Informazione e Bioingegneria  
{giuseppe.denittis, francesco1.trovo}@polimi.it

September 30, 2016

## Abstract

The present survey aims at presenting the current machine learning techniques employed in security games domains. Specifically, we focused on papers and works developed by the Teamcore of University of Southern California, which deepened different directions in this field. After a brief introduction on Stackelberg Security Games (SSGs) and the poaching setting, the rest of the work presents how to model a boundedly rational attacker taking into account her human behavior, then describes how to face the problem of having attacker's payoffs not defined and how to estimate them and, finally, presents how online learning techniques have been exploited to learn a model of the attacker.

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Stackelberg paradigm and SSG . . . . .	3
1.2	The poaching setting . . . . .	4
<b>2</b>	<b>Using human behavior models in solving SSGs</b>	<b>5</b>
2.1	SUQR: modeling a boundedly rational attacker . . . . .	5
2.1.1	Learning SUQR parameters . . . . .	6
2.2	From MATCH to SHARP . . . . .	7
2.2.1	Adaptive utility model . . . . .	8
2.2.2	SHARP's utility Computation . . . . .	9
2.2.3	Generating defender's strategies against SHARP . . . . .	9
2.3	Learning adversary models from three defender's strategies . . . . .	10
2.3.1	Linear utility functions . . . . .	11
2.3.2	Learning the optimal strategy . . . . .	12
<b>3</b>	<b>Determining attacker's payoffs exploiting regret-based solutions</b>	<b>13</b>
<b>4</b>	<b>Online learning</b>	<b>16</b>
4.1	Handling exploration-exploitation tradeoffs in Security Games . . . . .	16
4.1.1	Problem formulation . . . . .	17
4.1.2	Learning model from defenders' previous observations . . . . .	18
4.1.3	Restless multi-armed bandit problems . . . . .	18
4.1.4	Restless bandit formulation . . . . .	19
4.1.5	Sufficient condition for indexability . . . . .	20
4.1.6	Numerical evaluation of indexability . . . . .	21
4.1.7	Computation of Whittle index Policy . . . . .	21
4.1.8	Special POMDP formulation . . . . .	22
4.1.9	Value iteration for the special POMDP . . . . .	23
4.1.10	Planning from POMDP view . . . . .	24
4.2	Online planning for optimal protector strategies in resource conservation games . . . . .	24
4.2.1	Problem formulation . . . . .	25
4.2.2	GMOP algorithm . . . . .	27
4.2.3	Applying Gibbs sampling in GMOP . . . . .	28
	<b>References</b>	<b>30</b>

# 1 Introduction

The present survey aims at presenting the current main machine learning techniques employed in security games domains. Specifically, we focused on papers and works developed by the Teamcore of University of Southern California, which deepened different directions in this field. Among several works on this topic, e.g., [3, 9, 22, 15, 10, 11], this paper is essentially based on these works [12, 6, 4, 8, 17, 16].

After a brief introduction on Stackelberg Security Games (SSGs) and the poaching setting, the rest of the work is organized according to the different problems that has been dealt with.

- Section 2 shows how to model a boundedly rational attacker taking into account her human behavior.
- Section 3 faces the problem of having attacker's payoffs not defined and how to estimate them studying the regret of the defender.
- Section 4 presents how online learning techniques have been exploited to learn a model of the attacker.

## 1.1 Stackelberg paradigm and SSG

Usually, to represent security scenarios, a specific class of games is adopted, i.e., Stackelberg Games [23]. Here, on one side there is the Defender, which publicly commits to a mixed strategy, i.e., a probability distribution on the actions available to the player, and on the other side there is an Attacker, that observes the commitment of the Defender and acts consequently. Such games are called Stackelberg Security Games (SSG). Specifically, in SSGs, the Defender attempts to protect a set of  $T$  targets from an Attacker, by optimally allocating a set of  $R$  resources,  $R < T$ . Denote by  $\mathbf{x} = \{x_t\}$  the Defender's strategy where  $x_t$  is the coverage probability at target  $t$ , the set of feasible strategies is:  $\mathbf{X} = \{\mathbf{x} : 0 \leq x_t \leq 1, \sum_t x_t \leq R\}$ . If the adversary attacks  $t$  when the defender is not protecting it, the adversary receives a reward  $R_t^a$ , otherwise the adversary gets a penalty  $P_t^a$ . Conversely, the Defender receives a penalty  $P_t^d$  in the former case and a reward  $R_t^d$  in the latter case. Let  $(\mathbf{R}^a, \mathbf{P}^a)$  and  $(\mathbf{R}^d, \mathbf{P}^d)$  be the payoff vectors. The players' expected utilities at  $t$  is computed as:

$$\begin{aligned} U_t^a(x, R^a, P^a) &= x_t P_t^a + (1 - x_t) R_t^a \\ U_t^d(x, R^d, P^d) &= x_t R_t^d + (1 - x_t) P_t^d \end{aligned}$$

In general, in such situations, the most appropriate solution concept is the Leader-follower equilibrium. The problem of finding such equilibrium can be

formulated as:

$$\begin{aligned}
& \arg \max_{x_l, x_f^*} \sum_{a_l \in A_l} \sum_{a_f \in A_f} [U_l(a_l, a_f) x_l(a_l) x_f^*(a_f)] \\
& \text{s.t.} \quad \sum_{a_l \in A_l} x_l(a_l) = 1 \\
& \quad \quad x_l(a_l) \geq 0 \quad \quad \quad \forall a_l \in A_l \\
& x_f^* \in \arg \max_{x_f} \sum_{a_l \in A_l} \sum_{a_f \in A_f} [U_f(a_f, a_l) x_f(a_f) x_l(a_l)] \\
& \text{s.t.} \quad \sum_{a_f \in A_f} x_f(a_f) = 1 \\
& \quad \quad x_f(a_f) \geq 0 \quad \quad \quad \forall a_f \in A_f
\end{aligned}$$

where  $x_l(a_l)$  ( $x_f(a_f)$ ) is the probability that the leader (follower) will play action  $a_l$  ( $a_f$ ) and  $x_f^*$  is the best strategy of the follower.

In zero-sum games, the Leader-follower equilibrium coincides with the Nash equilibrium [21] and the maxmin/minmax strategies.

## 1.2 The poaching setting

Poaching and illegal over-fishing are critical international problems leading to destruction of ecosystems. For example, three out of nine tiger species have gone extinct in the past 100 years and others are now endangered due to poaching [18]. Law enforcement agencies in many countries are hence challenged with applying their limited resources to protecting endangered animals and fish stocks.

Building upon the success of applying SSGs to protect infrastructure including airports [14], ports [19] and trains [27], researchers are now applying game theory to green security domains, e.g., protecting fisheries from over-fishing [1, 5] and protecting wildlife from poaching [25]. There are several key features in green security domains that introduce novel research challenges.

1. The defender is faced with multiple adversaries who carry out repeated and frequent illegal activities (attacks), yielding a need to go beyond the one-shot SSG model.
2. In carrying out such frequent attacks, the attackers generally do not conduct extensive surveillance before performing an attack and spend less time and effort in each attack, and thus it becomes more important to model the attackers' bounded rationality and bounded surveillance.
3. There is more attack data available in green security domains than in infrastructure security domains, which makes it possible to learn the attackers' decision making model from data.

## 2 Using human behavior models in solving SSGs

In game theory, the adversary is usually represented as a fully rational player. In real-world, people are not fully rational, i.e., their choices are not simply determined by mere calculations. The problem that is addressed here is the representation of a boundedly rational attacker in SSGs. In fact, the fully rationality condition is relaxed to make an important step towards modeling real-world attacker, both to extract useful information about already gathered data, but also to build new models of attackers to enhance the current level of security.

### 2.1 SUQR: modeling a boundedly rational attacker

In SSGs, attacker bounded rationality is often modeled via behavior models such as Quantal Response (QR) [7]. The QR model predicts a stochastic distribution of the adversary response: the greater the expected value of a target the more likely the adversary will attack that target. QR’s key parameter  $\lambda$  represents the level of rationality in adversary’s response: as  $\lambda$  increases, the predicted response by the QR model converges to the optimal action of the adversary. Instead of using a human behavior model, MATCH, the best algorithm up to 2013, computes a robust defender strategy by guaranteeing a bound on the defender’s loss in her expected value if the adversary deviates from her optimal choice. More specifically, the defender’s loss is constrained to be no more than a factor of  $\beta$  times the adversary’s loss in his expected value. The key parameter  $\beta$  describes how much the defender is willing to sacrifice when the adversary deviates from the optimal action.

The key idea in Subjective Expected Utility (SEU) is that individuals have their own evaluations of each alternative during decision making. Recall that in an SSG, the information presented to the human subject for each choice includes: the marginal coverage on target  $t(x_t)$ ; the subject’s reward and penalty ( $R_t^a, P_t^a$ ); the defender’s reward and penalty ( $R_t^d, P_t^d$ ). Inspired by the idea of SEU, it has been proposed [12] a subjective utility function of the adversary for SSG as the following:

$$\hat{U}_t^a(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a) = w_1 x_t + w_2 R_t^a + w_3 P_t^a$$

In fact, SUQR is motivated by the lens model which suggested that evaluation of adversaries over targets is based on a linear combination of multiple observable features. One key advantage of these behavioral models is that they can be used to predict attack frequency for multiple attacks by the adversary, where in the attacking probability is a normalization of attack frequency.

The novelty of this subjective utility function is the linear combination of the values (rewards/penalty) and coverage probabilities. (Note that the decision-making of the general population is modeled, not of each individual since there are no sufficient data for each specific subject). This model actually leads to higher prediction accuracy than the classic expected value function. Other

alternatives to this subjective utility function are feasible, e.g., including all the information presented to the subjects:

$$\hat{U}_t^a(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a) = w_1 x_t + w_2 R_t^a + w_3 P_t^a + w_4 R_t^d + w_5 P_t^d$$

Then, the QR model is modified by replacing the classic expected value function with the SU function, leading to the SUQR model. In the SUQR model, the probability that the adversary chooses target  $t$ ,  $q_t$ , is given by:

$$q_t(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a) = \frac{e^{\lambda \hat{U}_t^a(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a)}}{\sum_{t'} e^{\lambda \hat{U}_{t'}^a(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a)}}$$

The problem of finding the optimal strategy for the defender can therefore be formulated as:

$$\begin{aligned} & \max_x \sum_{t=1}^T \frac{e^{\lambda U_t^a(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a)}}{\sum_{t'} e^{\lambda U_{t'}^a(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a)}} (x_t R_t^d + (1 - x_t) P_t^d) \\ & \text{s.t. } \sum_{t=1}^T x_t \leq K, 0 \leq x_t \leq 1 \end{aligned}$$

Here, the objective is to maximize the defender's expected value given that the adversary chooses to attack each target with a probability according to the SUQR model.

### 2.1.1 Learning SUQR parameters

Without loss of generality,  $\lambda = 1$ . As customarily done in traditional machine learning, Maximum Likelihood Estimation (MLE) to learn the parameters  $(w_1, w_2, w_3)$  is employed. Given the defender strategy  $\mathbf{x}$  and  $N$  samples of the players' choices, the log-likelihood of  $(w_1, w_2, w_3)$  is given by:

$$\log L(w_1, w_2, w_3 | \mathbf{x}) = \sum_{j=1}^N \log[q_{t_j}(w_1, w_2, w_3 | \mathbf{x})]$$

where  $t_j$  is the target that is chosen in sample  $j$  and  $q_{t_j}(w_1, w_2, w_3 | \mathbf{x})$  is the probability that the adversary chooses the target  $t_j$ . Let  $N_t$  be the number of subjects attacking target  $t$ . Then:

$$\log L(w_1, w_2, w_3 | \mathbf{x}) = \sum_{t=1}^T N_t \log[q_{t_j}(w_1, w_2, w_3 | \mathbf{x})]$$

It can be shown that the Hessian matrix of  $\log L[q_{t_j}(w_1, w_2, w_3 | \mathbf{x})]$  is negative semi-definite. Thus, this function has an unique local maximum point and hence a convex optimization solver can be used to compute the optimal weights  $(w_1, w_2, w_3)$ , e.g., `fmincon` in `Matlab`.

## 2.2 From MATCH to SHARP

Starting from the above results, a new model is now introduced, SHARP [6], which:

- reasons based on success or failure of the adversary’s past actions on exposed portions of the attack surface to model adversary adaptiveness;
- reasons about similarity between exposed and unexposed areas of the attack surface, and also incorporates a discounting parameter to mitigate adversary’s lack of exposure to enough of the attack surface;
- integrates a non-linear probability weighting function to capture the adversary’s true weighting of probability.

Following the approach presented in the previous section, MLE has been applied to learn the weights of the SUQR model based on data collected from our human subject experiments and found that the weights on coverage probability were positive for all the experiments. That is, counter-intuitively humans were modeled as being attracted to regions with high coverage probability, even though they were not attacking targets with very high coverage but they were going after targets with moderate to very low coverage probability.

To address this issue, a solution can be the augmentation of the Subjective Utility function with a two-parameter probability weighting function, that can be either inverse S-shaped (concave near probability zero and convex near probability one) or S-shaped.

$$f(p) = \frac{\delta p^\gamma}{\delta p^\gamma + (1 - p)^\gamma} \quad (1)$$

The SU of an adversary denoted by  $a$  can then be computed as:

$$SU_i^a(x) = w_1 f(x_i) + w_2 R_i^a + w_3 P_i^a,$$

where  $f(x_i)$  for coverage probability  $x_i$  is computed as per Equation (1). The two parameters  $\delta$  and  $\gamma$  control the elevation and curvature of the function respectively.  $\gamma < 1$  results in an inverse S-shaped curve while  $\gamma > 1$  results in an S-shaped curve. This is the PSU (Probability weighted Subjective Utility) function. The curve representing human weights for probability is S-shaped in nature, and not inverse S-shaped as prospect theory suggests. The S-shaped curve indicates that people would overweigh high probabilities and underweigh low to medium probabilities.

W.r.t. MATCH, SHARP introduces a new feature – distance – that affects the reward and hence the obvious question for us was to investigate the effect of this new feature in predicting adversary behavior. Several variations of PSU with different combinations of features can be considered.

$$SU_i^a(x) = w_1 f(x_i) + w_2 \phi_i + w_3 P_i^a \quad (2)$$

$$SU_i^a(x) = w_1 f(x_i) + w_2 R_i^a + w_3 P_i^a + w_4 D_i \quad (3)$$

$$SU_i^a(x) = w_1 f(x_i) + w_2 \phi_i + w_3 P_i^a + w_4 D_i \quad (4)$$

where  $\phi_i$  and  $D_i$  refer to the animal density at target  $i$  and the distance to target  $i$  from the poacher's starting location respectively.

### 2.2.1 Adaptive utility model

A second major innovation in SHARP is the adaptive nature of the adversary and addressing the issue of attack surface exposure. The attack surface  $\alpha$  is defined as the  $n$ -dimensional space of the features used to model adversary behavior.

For example, as per the third PSU model in Equation (4), this would mean the space represented by the following four features: coverage probability, animal density, adversary penalty and distance from the starting location.

A target profile  $\beta_k \in \alpha$  is defined as a point on the attack surface  $\alpha$  and can be associated with a target. Exposing the adversary to a lot of different target profiles would therefore mean exposing the adversary to more of the attack surface and gathering valuable information about their behavior. While a particular target location, defined as a distinct region in the 2-d space, can only be associated with one target profile in a particular round, more than one target may be associated with the same target profile in the same round.

**Observation 1.** *Adversaries who have succeeded in attacking a target associated with a particular target profile in one round, tend to attack a target with similar target profiles in next round.*

**Observation 2.** *Adversaries who have failed in attacking a target associated with a particular target profile in one round, tend not to attack a target with similar target profiles in the next round.*

The vulnerability associated with a target profile  $\beta_i$ , which was shown to the adversary in round  $r$ , denoted  $V_{\beta_i}^r$ , is defined as a function of the total number of successes and failures on the concerned target profile in that round (denoted by  $success_{\beta_i}^r$  and  $failure_{\beta_i}^r$  respectively).

$$V_{\beta_i}^r = \frac{success_{\beta_i}^r - failure_{\beta_i}^r}{success_{\beta_i}^r + failure_{\beta_i}^r}$$

Therefore, more successful attacks and few failures on a target profile indicate that it was highly vulnerable in that round. Because multiple targets can be associated with the same target profile and the pure strategy generated based on the mixed strategy  $x$  in a particular round may result in a defender



being present at some of these targets while not at others, there may be both successes and failures associated with the same target profile in that round.

The attractiveness of a target profile  $\beta_i$  at the end of round  $R$ , denoted  $A_{\beta_i}^R$ , is defined as a function of the vulnerabilities for  $\beta_i$  from round 1 to round  $R$ .

$$A_{\beta_i}^R = \frac{\sum_{r=1}^R V_{\beta_i}^r}{R}$$

Therefore, the attractiveness of a target profile is modeled as the average of the vulnerabilities for that target profile over all the rounds till round  $R$ . This is consistent with the notion that a target profile which has led to more successful attacks over several rounds will be perceived as more attractive by the adversary.

### 2.2.2 SHARP's utility Computation

Existing models (such as SUQR) only consider the adversary's actions from round  $(r - 1)$  to predict their actions in round  $r$ . However, it is clear that the adversary's actions in a particular round are dependent on his past successes and failures. Thus, a novel adaptive probability weighted subjective utility function that captures this adaptive nature of the adversary's behavior by capturing the shifting trends in attractiveness of various target profiles over rounds is proposed.

$$\begin{aligned} ASU_{\beta_i}^R = & (1 - d * A_{\beta_i}^R)w_1f(x_{\beta_i}) + (1 + d * A_{\beta_i}^R)w_2\phi_{\beta_i} + \\ & + (1 + d * A_{\beta_i}^R)w_3P_{\beta_i}^a + (1 - d * A_{\beta_i}^R)w_4D_{\beta_i} \end{aligned}$$

$d$  is a discounting parameter which is based on a measure of the amount of attack surface exposed.  $d$  is low in the initial rounds when the defender does not have enough of the right kind of data, but would gradually increase as more information about the attacker's preferences about various regions of the attack surface become available.

Now, let us look at how reasoning about unexposed portions of the attack surface based on the exposed areas. If a target profile  $\beta_u$  was not exposed to attacker response in round  $r$ , the defender will not be able to compute the vulnerability  $V_{\beta_u}^r$ . Therefore, it is not possible to estimate the attractiveness for  $\beta_u$  and hence the optimal defender strategy. So, in keeping with the analysis on available data and based on the spillover effect introduced earlier, the distance-weighted  $k$ -nearest neighbors algorithm is employed to obtain the vulnerability  $V_{\beta_u}^r$  of an unexposed target profile  $\beta_u$  in round  $r$ , based on the  $k$  most similar target profiles which were exposed to the attacker in round  $r$ .

### 2.2.3 Generating defender's strategies against SHARP

While SHARP provides an adversary model, the defender strategies against such model must be generated. To that end, first the parameters of SHARP from

available data are learned. Then, future round strategies against the boundedly rational adversary are generated, characterized by the learned model parameters by solving the following optimization problem:

$$\max_{x \in X} \left[ \sum_{i \in T} U_i^d(x) q_i^R(w|x) \right]$$

where  $q_i^R(w|x)$  is the probability that the adversary will attack target  $i$  in round  $R$ .

### 2.3 Learning adversary models from three defender's strategies

Here, a new approach to learn the parameters of the behavioral model of a bounded rational attacker (thereby pinpointing a near optimal strategy) is developed, by observing how the attacker responds to only three defender strategies.

*Note: even though the setting is the same of the previous sections, since some assumptions changed, e.g., the adaptivity of the attacker, we slightly change the adopted notation to avoid confusion.*

Let the utility function of the attacker for target  $t \in T$  be denoted by  $u_t : [0, 1] \rightarrow R$ . Given a coverage probability vector  $\mathbf{x} \in \mathbf{X}$ , the utility of the attacker under strategy  $\mathbf{x}$  is defined as  $u_t(x_t)$ . Upon observing the defender's strategy  $\mathbf{x}$ , the attacker computes the utility on each target  $t$ ,  $u_t(x_t)$ , and based on these utilities responds to the defender's strategy. Here, a non-adaptive attacker is considered. She attacks target  $t$  with probability:

$$D^{\mathbf{x}}(t) = \frac{e^{u_t(x_t)}}{\sum_{i \in T} e^{u_i(x_i)}}$$

Our model is a generalization of bounded rationality models such as SUQR. Suppose the same mixed strategy  $\mathbf{x}$  is played for multiple time steps. The empirical distribution of attacks on target  $t$  under  $\mathbf{x}$  is denoted by  $\hat{D}^{\mathbf{x}}(\cdot)$ . Furthermore, it is assumed that for the strategies, and for all  $t$ ,  $D^{\mathbf{x}}(t) \geq \rho$ , for some  $\rho = \frac{1}{\text{poly}(n)}$ . This assumption is required to estimate the value of  $D^{\mathbf{x}}(t)$  with polynomially many samples. Our goal is to learn the utility functions,  $u_t(\cdot)$  for all  $t \in T$ , by observing attacker's responses to a choice of coverage probability vectors  $\mathbf{x} \in \mathbf{X}$ . This allows to find an approximately optimal defender strategy — the strategy that leads to the best defender utility.  $\hat{u}_t : [0, 1] \rightarrow R$  uniformly approximates or uniformly learns  $u_t(\cdot)$  within an error of  $\epsilon$ , if  $\forall x \in [0, 1], |\hat{u}_t(x) - u_t(x)| \leq \epsilon$ . Note that the attacker's mixed strategy remains the same when the utility functions corresponding to all targets are increased by the same value. Therefore, only a normalized representation of the utility functions can be learned, such that for all  $t$  and all  $x$ ,  $|\hat{u}_t(x) + c - u_t(x)| \leq \epsilon$  for some  $c$ .

### 2.3.1 Linear utility functions

Assume that the utility functions are linear and denoted by  $u_t(x) = w_t x + c_t$ . Normalizing the utilities, without loss of generality,  $c_n = 0$ .

**Theorem 1.** *Suppose the functions  $u_1(\cdot), \dots, u_n(\cdot)$  are linear. Consider any 3 strategies  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{X}$  such that for any  $t < n$ ,  $|(x_t - y_t)(x_n - z_n) - (x_n - y_n)(x_t - z_t)| \geq \lambda$ , and for any two different strategies  $\mathbf{p}, \mathbf{q} \in \{\mathbf{x}, \mathbf{y}, \mathbf{z}\}$ , it holds  $|p_t - q_t| \geq \nu$ . If it is possible to access  $m = \Omega\left(\frac{1}{\rho} \left(\frac{1}{\epsilon\nu\lambda}\right)^2 \log\left(\frac{n}{\delta}\right)\right)$  samples of each of these strategies, then with probability  $1 - \delta$ , each  $u_t(\cdot)$  can be uniformly learned within error  $\epsilon$ .*

$\nu$  depends on how different the strategies are from each other — a very small value means that they are almost identical on some coordinates. The lower bound of  $\lambda$  would not be very small unless there is a very specific relation between the strategies. As a sanity check, if the three strategies were chosen uniformly at random from the simplex, both values would be at least  $1/\text{poly}(n)$ . In the quantal best-response model, for each strategy  $\mathbf{x}$ , the ratio between the attack probabilities of two targets  $t$  and  $n$  follows the relation:

$$u_t(x_t) = \ln\left(\frac{D^{\mathbf{x}}(t)}{D^{\mathbf{x}}(n)}\right)$$

Therefore, each strategy induces  $n - 1$  linear equations that can be used to solve for the coefficients of  $u_t$ . However, only an estimate  $\hat{D}^{\mathbf{x}}(t)$  of the probability that target  $t$  is attacked under a strategy  $\mathbf{x}$ , based on the given samples, can be obtained. So, the inaccuracy in our estimates of  $\ln\left(\frac{\hat{D}^{\mathbf{x}}(t)}{\hat{D}^{\mathbf{x}}(n)}\right)$  leads to inaccuracy in the estimated polynomial  $\hat{u}_t$ . For sufficiently accurate estimates  $D^{\mathbf{x}}(t)$ , the value of  $u_t$  differs from the true value by at most  $\epsilon$ .

**Lemma 1.** *Given  $\mathbf{x} \in \mathbf{X}$ , let  $D^{\mathbf{x}}(t)$  be the empirical distribution of attacks based on  $m = \Omega\left(\frac{1}{\rho} \left(\frac{1}{\epsilon\nu\lambda}\right)^2 \log\left(\frac{n}{\delta}\right)\right)$  samples. With probability  $1 - \delta$ , for all  $t \in T$ ,  $\frac{1}{\epsilon} \leq \frac{\hat{D}^{\mathbf{x}}(t)}{D^{\mathbf{x}}(t)} \leq 1 + \epsilon$ .*

**Theorem 2.** *Suppose the functions  $u_1(\cdot), \dots, u_n(\cdot)$  are polynomials of degree at most  $d$ . Consider any  $2d+1$  strategies,  $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(d)}, \mathbf{y}^{(d+1)} = \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(d)} \mathbf{x}^{(d+1)}$  such that for all  $k, k', k \neq k'$ ,  $y_1^{(k)} = y_1^{(k')}$ ,  $p_n^{(k)} = p_n^{(k')}$ ,  $|y_n^{(k)} - y_n^{(k')}| \geq \nu$  and for all  $t < n$ ,  $|x_t^{(k)} - x_t^{(k')}| \geq \nu$ . If it is possible to have access to  $m = \Omega\left(\frac{1}{\rho} \left(\frac{1}{\epsilon\nu\lambda}\right)^2 \log\left(\frac{n}{\delta}\right)\right)$  samples of each of these strategies, then with probability  $1 - \delta$ , each  $u_t(\cdot)$  can be uniformly learned within error  $\epsilon$ .*

Now, any utility function that is continuous and  $L$ -Lipschitz, i.e., for all  $t$  and values  $x$  and  $y$ ,  $|u_t(x) - u_t(y)| \leq L|x - y|$ , should be learned. Such utility functions can be uniformly learned up to error  $\epsilon$ , using  $O\left(\frac{L}{\epsilon}\right)$  strategies. For any  $L$ -Lipschitz function  $u_t(x)$ , there is a polynomial of degree  $m = 12L/\epsilon$  that uniformly approximates  $u_t(x)$  within error of  $\epsilon/2$ .

**Theorem 3.** Suppose the functions  $u_1(\cdot), \dots, u_n(\cdot)$  are  $L$ -Lipschitz. For  $d = 12L/\epsilon$ , consider any  $2d+1$  strategies,  $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(d)}, \mathbf{y}^{(d+1)} = \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(d)} \mathbf{x}^{(d+1)}$  such that for all  $k, k', k \neq k', y_1^{(k)} = y_1^{(k')}, p_n^{(k)} = p_n^{(k')}, |y_n^{(k)} - y_n^{(k')}| \geq \nu$  and for all  $t < n, |x_t^{(k)} - x_t^{(k')}| \geq \nu$ . If it is possible to have access to  $m = \Omega\left(\frac{L^2}{\rho \epsilon^{24L/\epsilon}} \log\left(\frac{n}{\delta}\right)\right)$  samples of each of these strategies, then with probability  $1 - \delta$ , each  $u_t(\cdot)$  can be uniformly learned within error  $\epsilon$ .

### 2.3.2 Learning the optimal strategy

So far, the focus has been on the problem of uniformly learning the utility function of the attacker. Now, it is shown that an accurate estimate of this utility function allows to pinpoint an almost optimal strategy for the defender. Let the utility function of the defender on target  $t \in T$  be denoted by  $v_t : [0, 1] \rightarrow [-1, 1]$ . Given a coverage probability vector  $\mathbf{x} \in \mathbf{X}$ , the utility the defender receives when target  $t$  is attacked is  $v_t(x_t)$ . The overall expected utility of the defender is:

$$V(\mathbf{x}) = \sum_{t \in T} D^{\mathbf{x}}(t) v_t(x_t)$$

Let  $\hat{u}_t$  be the learned attacker utility functions, and  $\bar{D}^{\mathbf{x}}(t)$  be the predicted attack probability on target  $t$  under strategy  $\mathbf{x}$ , according to the utilities  $\hat{u}_t$ , i.e.:

$$\bar{D}^{\mathbf{x}}(t) = \frac{e^{\hat{u}_t(x_t)}}{\sum_{i \in T} e^{\hat{u}_i(x_i)}}$$

Let  $\bar{V}(\mathbf{x})$  be the predicted expected utility of the defender based on the learned attacker utilities  $\bar{D}^{\mathbf{x}}(t)$ , i.e.,  $\bar{V}(\mathbf{x}) = \sum_{t \in T} \bar{D}^{\mathbf{x}}(t) v_t(x_t)$ . When the attacker utilities are uniformly learned within error  $\epsilon$ , then  $\bar{V}$  estimates  $V$  with error at most  $8\epsilon$ . At a high level, this is established by showing that one can predict the attack distribution using the learned attacker utilities. Furthermore, optimizing the defender's strategy against the approximate attack distributions leads to an approximately optimal strategy for the defender.

**Theorem 4.** Assume for all  $\mathbf{x}$  and any  $t \in T$ ,  $|\hat{u}_t(x_t) - u_t(x_t)| \leq \epsilon \leq 1/4$ . Then, for all  $\mathbf{x}$ ,  $|\bar{V}(\mathbf{x}) - V(\mathbf{x})| \leq 4\epsilon$ . Furthermore, let  $\mathbf{x}' = \arg \max_{\mathbf{x}} \bar{V}(\mathbf{x})$  be the predicted optimal strategy, then  $\max_{\mathbf{x}} V(\mathbf{x}) - V(\mathbf{x}') \leq 8\epsilon$ .

### 3 Determining attacker's payoffs exploiting regret-based solutions

The adversary behavior models (capturing bounded rationality) can be learned from real-world data on where adversaries have attacked, and game payoffs can be determined precisely from data on animal densities.

One key approach to modeling payoff uncertainty is to express the adversary's payoffs as lying within specific intervals: for each target  $t$ , it holds:

$$R_t^a \in [R_{min}^a(t), R_{max}^a(t)], P_t^a \in [P_{min}^a(t), P_{max}^a(t)]$$

While a behavioral model from real-world data can be learned, access to data to precisely compute animal density may be not always possible. For example, given limited numbers of rangers, they may have patrolled and collected wildlife data from only a small portion of a national park, and thus payoffs in other areas of the park may remain uncertain. Also, due to the dynamic changes (e.g., animal migration), players' payoffs may become uncertain in the next season. Hence, a new MiniMaxRegret (MMR)-based robust algorithm is introduced, ARROW, to handle payoff uncertainty in Green Security Games (a.k.a. GSGs, i.e., Security Games for the protection of wildlife, forest and fisheries), taking into account adversary behavioral models. Here, the main focus is on zero-sum games, as motivated by recent work in green security domains and earlier major SSG applications that use zero-sum games. In addition, a model inspired by SUQR is adopted, where the subjective utility function is:

$$\hat{U}_t^a(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a) = w_1 x_t + w_2 R_t^a + w_3 P_t^a + w_4 \Phi_t$$

where  $x_t$  is the coverage probability of target  $t$  and  $\Phi_t$  is another feature of target  $t$ , e.g., distance, as seen for SHARP. MMRb with uncertain payoffs is now formulated for both players in zero-sum SSG with a boundedly rational attacker.

**Definition 1.** *Given  $(\mathbf{R}^a, \mathbf{P}^a)$ , the defender's behavioral regret is the loss in her utility for playing a strategy  $\mathbf{x}$  instead of the optimal strategy, which is represented as follows:*

$$R_b(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a) = \max_{\mathbf{x}' \in X} F(\mathbf{x}', \mathbf{R}^a, \mathbf{P}^a) - F(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a)$$

where

$$F(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a) = \sum_t \hat{q}_t(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a) U_t^d(\mathbf{x}, \mathbf{R}^d, \mathbf{P}^d)$$

Behavioral regret measures the distance in terms of utility loss from the defender strategy  $\mathbf{x}$  to the optimal strategy given the attacker payoffs. Here,  $F(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a)$  is the defender's utility for playing  $\mathbf{x}$  where the attacker payoffs, whose response follows SUQR, are  $(\mathbf{R}^a, \mathbf{P}^a)$ . The defender's payoffs in zero-sum games are  $\mathbf{R}^d = -\mathbf{P}^a$  and  $\mathbf{P}^d = -\mathbf{R}^a$ . When the payoffs are uncertain,

if the defender plays a strategy  $\mathbf{x}$ , she receives different behavioral regrets w.r.t to different payoff instances within the uncertainty intervals. Thus, she could receive a behavioral max regret which is defined as follows.

**Definition 2.** *Given payoff intervals  $I$ , the behavioral max regret for the defender to play a strategy  $\mathbf{x}$  is the maximum behavioral regret over all payoff instances:*

$$MR_b(\mathbf{x}, I) = \max_{(\mathbf{R}^a, \mathbf{P}^a) \in I} R_b(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a)$$

**Definition 3.** *Given payoff intervals  $I$ , the behavioral minimax regret problem attempts to find the defender optimal strategy that minimizes the MRb she receives:*

$$MMR_b(I) = \min_{\mathbf{x} \in X} MR_b(\mathbf{x}, I)$$

Intuitively, behavioral minimax regret ensures that the defender's strategy minimizes the loss in the solution quality over the uncertainty of all possible payoff realizations.

Overall, MMRb can be reformulated as minimizing the max regret  $r$  such that  $r$  is no less than the behavioral regrets over all payoff instances within the intervals:

$$\begin{aligned} \min_{\mathbf{x} \in X, r \in \mathbb{R}} \quad & r \\ \text{s.t.} \quad & r \geq F(\mathbf{x}', \mathbf{R}^a, \mathbf{P}^a) - F(\mathbf{x}, \mathbf{R}^a, \mathbf{P}^a), \forall (\mathbf{R}^a, \mathbf{P}^a) \in I, \mathbf{x}' \in X \end{aligned}$$

The set of constraints is infinite since  $X$  and  $I$  are continuous. One practical approach to optimization with large constraint sets is constraint sampling, coupled with constraint generation (a.k.a. row generation). Following this approach, ARROW samples a subset of constraints and gradually expands this set by adding violated constraints to the relaxed problem until convergence to the optimal  $MMR_b$  solution.

Specifically, ARROW begins by sampling pairs  $(\mathbf{R}^a, \mathbf{P}^a)$  of the adversary payoffs uniformly from  $I$ . The corresponding optimal strategies for the defender given these payoff samples, denoted  $\mathbf{x}'$ , are then computed using the PASAQ algorithm [26] to obtain a finite set  $S$  of sampled constraints. These sampled constraints are then used to solve the corresponding relaxed MMRb program using the R.ARROW algorithm. The optimal solution  $(lb, \mathbf{x}^*)$  is thus obtained, providing a lower bound ( $lb$ ) on the true  $MMR_b$ . Then constraint generation is applied to determine violated constraints (if any). This uses the M.ARROW algorithm which computes  $MR_b(\mathbf{x}^*, I)$ , the optimal regret of  $\mathbf{x}^*$  which is an upper bound ( $ub$ ) on the true MMRb. If  $ub > lb$ , the optimal solution of M.ARROW,  $\{\mathbf{x}'^*, \mathbf{R}^{a,*}, \mathbf{P}^{a,*}\}$  provides the maximally violated constraint, which is added to  $S$ . Otherwise,  $\mathbf{x}^*$  is the minimax optimal strategy and  $lb = ub = MMR_b(I)$ .

The first step of ARROW is to solve the relaxed  $MMR_b$  problem using R.ARROW. This relaxed  $MMR_b$  problem is non-convex. Thus, R.ARROW presents two key ideas for efficiency:

1. binary search (which iteratively searches the defender’s utility space to find the optimal solution) to remove the fractional terms in relaxed  $MMR_b$ ;
2. it then applies piecewise-linear approximation to linearize the non-convex terms of the resulting decision problem at each binary search step.

**Theorem 5.** *R.ARROW provides an  $O(\epsilon + \frac{1}{M})$ -optimal solution of relaxed  $MMR_b$  where  $\epsilon$  is the tolerance of binary search and  $M$  is the number of piece-wise segments.*

Given the optimal solution  $\mathbf{x}^*$  returned by R.ARROW, the second step of ARROW is to compute  $MR_b$  of  $\mathbf{x}^*$  using M.ARROW. A local search with multiple starting points is employed since allows to reach different local optima.

While ARROW incorporates an adversary behavioral model, it may not be applicable for green security domains where there may be a further paucity of data in which not only payoffs are uncertain but also parameters of the behavioral model are difficult to learn accurately. Therefore, a novel MMR-based algorithm, ARROW-Perfect, can be introduced to handle uncertainty in both players’ payoffs assuming a perfectly rational attacker. In general, ARROW-Perfect follows the same constraint sampling and constraint generation methodology as ARROW. Yet, by leveraging the property that the attacker’s optimal response is a pure strategy (given a perfectly rational attacker) and the game is zero-sum, the exact optimal solutions for computing both relaxed MMR and max regret in polynomial time are obtained (while we cannot provide such guarantees for a boundedly rational attacker).

## 4 Online learning

### 4.1 Handling exploration–exploitation tradeoffs in Security Games

Previous research optimizes defenders’ strategies by modeling this problem as a repeated Stackelberg game, capturing the special property in this domain frequent interactions between defenders and attackers. However, this research fails to handle exploration–exploitation tradeoff in this domain caused by the fact that defenders only have knowledge of attack activities at targets they protect. The problem is formulated as a Restless Multi-Armed Bandit (RMAB) model to address this challenge. To use Whittle index policy to plan for patrol strategies in the RMAB, two sufficient conditions for indexability and an algorithm to numerically evaluate indexability are provided. Given indexability, a binary search based algorithm to find Whittle index policy efficiently is proposed.

It is assumed that defenders have knowledge of all poaching activities throughout the wildlife protected area. Unfortunately, given vast geographic areas for wildlife protection, defenders do not have knowledge of poaching activities in areas they do not protect. Thus, defenders are faced with the exploration–exploitation tradeoff. The exploration–exploitation tradeoff here is different from that in the non-Bayesian stochastic multi-armed bandit problem. In stochastic multi-armed bandit problems, the rewards of every arm are random variables with a stationary unknown distribution. However, in our problem, patrol affects attack activities, i.e.m, more patrol is likely to decrease attack activities and less patrol is likely to increase attack activities. Thus, the random variable distribution is changing depending on player’s choice — more selection (patrol) leads to lower reward (less attack activities) and less selection (patrol) leads to higher reward (more attack activities). On the other hand, adversarial multi-armed bandit problem is also not an appropriate model for this domain. In adversarial multi-armed bandit problems, the reward can arbitrarily change while the attack activities in our problem are unlikely to change rapidly in a short period.

Poaching activity is a dynamic process affected by patrol. If patrollers patrol in a certain location frequently, it is very likely that the poachers poaching in this location will switch to other locations for poaching. On the other hand, if a location has not been patrolled for a long time, poachers may gradually notice that and switch to this location for poaching. In the wildlife protection domain, both patrollers and poachers do not have perfect observation of their opponents’ actions. This observation imperfection lies in two aspects: limited observability — patrollers/poachers do not know what happens at locations they do not patrol/poach; partial observability — patrollers/poachers do not have perfect observation even at locations they patrol/poach — the location might be large (e.g., a  $2\text{km} \times 2\text{km}$  area) so that it is possible that patrollers and poachers do not see each other even if they are at the same location. These two properties make it extremely difficult for defenders to optimally plan their patrol strategies.



#### 4.1.1 Problem formulation

There are  $n$  targets that are indexed by  $N = \{1, \dots, n\}$ . Defenders have  $k$  patrol resources that can be deployed to these  $n$  targets. At every round, defenders choose  $k$  targets to protect. After that, defenders will have an observation of the number of attack activities for targets they protect, and no information for targets they do not protect. The objective for defenders is to decide which  $k$  targets to protect at every round to catch as many attackers as possible. Due to the partial observability on defenders' side (defenders' observation of attack activities is not perfect even for targets they protect), a hidden variable attack intensity is introduced, representing the true degree of attack intensity at a certain target. Clearly, this hidden variable attack intensity cannot directly be observed by defenders. Instead, defenders' observation is a random variable conditioned on this hidden variable attack intensity, and the larger the attack intensity is, the more likely it is for defenders to observe more attack activities during their patrol. The hidden variable attack intensity is discretized into  $n_s$  levels, denoted by  $S = \{0, 1, \dots, n_s - 1\}$ . Lower  $i$  represents lower attack intensity. For a certain target, its attack intensity transitions after every round. If this target is protected, attack intensity transitions according to a  $n_s \times n_s$  transition matrix  $T^1$ ; if this target is not protected, attack intensity transitions according to another  $n_s \times n_s$  transition matrix  $T^0$ . The transition matrix represents how patrol affects attack intensity —  $T^1$  tends to reduce attack intensity and  $T^0$  tends to increase attack intensity. Note that different targets may have different transition matrices because some targets may be more attractive to attackers (for example, some locations may have more animal resources in the wildlife protection domain) so that it is more difficult for attack intensity to go down and easier for attack intensity to go up. Also defenders' observations of attack activities are discretized into  $n_o$  levels, denoted by  $\mathbf{O} = \{0, 1, \dots, n_o - 1\}$ . Lower  $i$  represents less attack activities defenders observe. Note that defenders will only have observation for targets they protect. A  $n_s \times n_o$  observation matrix  $O$  determines how the observation depends on the hidden variable attack intensity. Generally, the larger the attack intensity is, the more likely it is for defenders to observe more attack activities during their patrol. Similar to transition matrices, different targets may have different observation matrices. While defenders get observations of attack activities during their patrol, they also receive rewards for that — arresting poachers/fareevaders/smugglers bring benefit. Clearly, the reward defenders receive depends on their observation and thus the reward function is defined as  $R(o), o \in \mathbf{O}$  — larger  $i$  leads to higher reward  $R(i)$ . Note that defenders only get rewards for targets they protect. To summarize, for the targets defenders protect, defenders get an observation depending on its current attack intensity, get the reward associated with the observation, and then the attack intensity transitions according to  $T^1$ ; for the targets defenders do not protect, defenders do not have any observation, get reward 0 and the attack intensity transitions according to  $T^0$ .

#### 4.1.2 Learning model from defenders' previous observations

Given defenders' action history  $\{a_i\}$  and observation history  $\{o_i\}$ , our objective is to learn the transition matrices  $T^1$  and  $T^0$ , observation matrix  $O$  and initial belief  $\pi$ . Due to the existence of hidden variables  $\{s_i\}$ , expectation-maximization (EM) algorithm is used for learning. The update steps are the following:

$$\begin{aligned}\pi_i^{(d+1)} &= P(s_1 = i|x; \theta^d) \\ T_{ij}^{1(d+1)} &= \frac{\sum_{t=1:T-1} P(s_t = i, s_{t+1} = j|x; \theta^d)}{\sum_{t=1:T-1} P(s_t = i|x; \theta^d)} \\ T_{ij}^{0(d+1)} &= \frac{\sum_{t=1:T-1} P(s_t = i, s_{t+1} = j|x; \theta^d)}{\sum_{t=1:T-1} P(s_t = i|x; \theta^d)} \\ O_{ij}^{(d+1)} &= \frac{\sum_{t=1:T} P(s_t = i|x; \theta^d) I(o_t = j)}{\sum_{t=1:T} P(s_t = i|x; \theta^d)}\end{aligned}$$

#### 4.1.3 Restless multi-armed bandit problems

In RMABs, each arm represents an independent Markov machine. At every round, the player chooses  $k$  out of  $n$  arms ( $k < n$ ) to activate and receives the reward determined by the state of the activated arms. After that, the states of all arms will transition to new states according to certain Markov transition probabilities. The problem is called *restless* because the states of passive arms will also transition like active arms. The aim of the player is to maximize his cumulative reward by choosing which arms to activate at every round. It is PSPACE-hard to find the optimal strategy to general RMABs [13]. An index policy assigns an index to each state of each arm to measure how rewarding it is to activate an arm at a particular state. At every round, the index policy chooses to pick the  $k$  arms whose current states have the highest indices. Since the index of an arm only depends on the properties of this arm, index policy reduces an  $n$ -dimensional problem to  $n$  1-dimensional problems so that the complexity is reduced from exponential with  $n$  to linear with  $n$ . Whittle proposed a heuristic index policy for RMABs by considering the Lagrangian relaxation of the problem [24]. It has been shown that Whittle index policy is asymptotically optimal under certain conditions as  $k$  and  $n$  tend to infinity with  $k/n$  fixed. When  $k$  and  $n$  are finite, extensive empirical studies have also demonstrated the near-optimal performance of Whittle index policy. Whittle index measures how attractive it is to activate an arm based on the concept of subsidy for passivity. It gives the subsidy  $m$  to passive action (not activate) and the smallest  $m$  that would make passive action optimal for the current state is defined to be the Whittle index for this arm at this state. Whittle index policy chooses to activate the  $k$  arms with the highest Whittle indices. Intuitively, the larger the  $m$  is, the larger the gap is between active action (activate) and

passive action, the more attractive it is for the player to activate this arm.

Mathematically, denote  $V_m(x; a = 0)$  ( $V_m(x; a = 1)$ ) to be the maximum cumulative reward the player can achieve until the end if he takes passive (active) action at the first round at the state  $x$  with subsidy  $m$ . Whittle index  $I(x)$  of state  $x$  is then defined to be:

$$I(x) = \inf_m \{m : V_m(x; a = 0) \geq V_m(x; a = 1)\}$$

However, Whittle index only exists and Whittle index policy can only be used when the problem satisfies a property known as indexability. Define  $\Phi(m)$  to be the set of states for which passive action is the optimal action given subsidy  $m$ :

$$\Phi(m) = \{x : V_m(x; a = 0) \geq V_m(x; a = 1)\}$$

**Definition 4.** *An arm is indexable if  $\Phi(m)$  monotonically increases from  $\emptyset$  to the whole state space as  $m$  increases from  $-\infty$  to  $+\infty$ . An RMAB is indexable if every arm is indexable.*

#### 4.1.4 Restless bandit formulation

Every target is viewed as an arm and defenders choose  $k$  arms to activate ( $k$  targets to protect) at every round. Consider a single arm (target), it is associated with  $n_s$  (hidden) states,  $n_o$  observations,  $n_s \times n_s$  transition matrices  $T^1$  and  $T^0$ ,  $n_s \times n_o$  observation matrix  $O$  and reward function  $R(o), o \in O$ . For the arm defenders activate, defenders get an observation, get reward associated with the observation, and the state transitions according to  $T^1$ . Note that defenders' observation is not the state. Instead, it is a random variable conditioned on the state, and reveals some information about the state. For the arms defenders do not activate, defenders do not have any observation, get reward 0 and the state transitions according to  $T^0$ . Since defenders can not directly observe the state, defenders maintain a belief  $b$  of the states for each target, based on which defenders make decisions. The belief is updated according to the Bayesian rules. The following equation shows the belief update when defenders protect this target ( $a = 1$ ) and get observation  $o$  or defenders do not protect this target ( $a = 0$ ).

$$b'(s') = \begin{cases} \eta \sum_{s \in \mathcal{S}} b(s) O_{so} T_{ss'}^1, & a = 1 \\ \sum_{s \in \mathcal{S}} b(s) O_{so} T_{ss'}^0, & a = 0 \end{cases}$$

where  $\eta$  is the normalization factor. When defenders do not protect this target ( $a = 0$ ), defenders do not have any observation, so their belief is updated according to the state transition rule. When defenders protect this target ( $a = 1$ ), their belief is firstly updated according to their observation  $o$  ( $b_{new}(s) = \eta b(s) O_{so}$  according to Bayes' rule), and then the new belief is then updated according to the state transition rule:

$$\begin{aligned}
b'(s') &= \sum_{s \in \mathcal{S}} b_{new}(s) T_{ss'}^1 \\
&= \sum_{s \in \mathcal{S}} \eta b(s) O_{so} T_{ss'}^1 \\
&= \eta \sum_{s \in \mathcal{S}} b(s) O_{so} T_{ss'}^1
\end{aligned}$$

Now the mathematical definition of Whittle index is presented. Denote  $V_m(b)$  to be the value function for belief state  $b$  with subsidy  $m$ .  $V_m(b; a = 0)$  to be the value function for belief state  $b$  with subsidy  $m$  and defenders take passive action.  $V_m(b; a = 1)$  to be the value function for belief state  $b$  with subsidy  $m$  and defenders take active action. The following equations show these value functions:

$$\begin{aligned}
V_m(b; a = 0) &= m + \beta V_m(b_{a=0}) \\
V_m(b; a = 1) &= \sum_{s \in \mathcal{S}} b(s) \sum_{o \in \mathcal{O}} O_{so} R(o) + \beta \sum_{o \in \mathcal{O}} \sum_{s \in \mathcal{S}} b(s) O_{so} V_m(b_{a=1}^o) \\
V_m(b) &= \max\{V_m(b; a = 0), V_m(b; a = 1)\}
\end{aligned}$$

When defenders take passive action, they get the immediate reward  $m$  and the  $\beta$ -discounted future reward — value function at new belief  $b_a = 0$ , which is updated from  $b$  according to the case  $a = 0$ . When defenders take active action, they get the expected immediate reward  $\sum_{s \in \mathcal{S}} b(s) \sum_{o \in \mathcal{O}} O_{so} R(o)$  and the  $\beta$ -discounted future reward. The value function  $V_m(b)$  is the maximum of  $V_m(b; a = 0)$  and  $V_m(b; a = 1)$ . Whittle index  $I(b)$  of belief state  $b$  is then defined to be:

$$I(b) = \inf_m \{m : V_m(b; a = 0) \geq V_m(b; a = 1)\}$$

The passive action set  $\Phi(m)$ , which is the set of belief states for which passive action is the optimal action given subsidy  $m$  is then defined to be:

$$\Phi(m) = \inf_m \{b : V_m(b; a = 0) \geq V_m(b; a = 1)\}$$

#### 4.1.5 Sufficient condition for indexability

Two sufficient conditions for indexability when  $n_o = 2$  and  $n_s = 2$  are provided. Denote the transition matrices to be  $T^0$  and  $T^1$ , observation matrix to be  $O$ . Clearly in our problem,  $O_{11} > O_{01}, O_{00} > O_{10}$  (higher attack intensity leads to higher probability to see attack activities when patrolling);  $T_{11}^1 > T_{01}^1, T_{00}^1 > T_{10}^1, T_{11}^0 > T_{01}^0, T_{00}^0 > T_{10}^0$  (positively correlated arms).

Define  $\alpha = \max\{T_{11}^0 - T_{01}^1, T_{11}^0 > T_{01}^1\}$ . Since it is a two-state problem with  $S = \{0, 1\}$ ,  $x$  represents the belief state:  $x = b(s = 1)$ , which is the probability of being in state 1.

Define  $\Gamma_1(x) = xT_{11}^1 + (1-x)T_{01}^1$  which is the belief for the next round if the belief for the current round is  $x$  and the active action is taken. Similarly,  $\Gamma_0(x) = xT_{11}^0 + (1-x)T_{01}^0$ , which is the belief for the next round if the belief for the current round is  $x$  and the passive action is taken. Below two theorems demonstrating two sufficient conditions for indexability are presented.

**Theorem 6.** *When  $\alpha \leq 0.5$ , the process is indexable, i.e., for any belief  $x$ , if  $V_m(x; a=0) \geq V_m(x; a=1)$ , then  $V_m(x; a=0) \geq V_m(x; a=1), \forall m \geq m$ .*

**Theorem 7.** *When  $\alpha\beta \leq 0.5$  and  $\Gamma_1(1) \leq \Gamma_0(0)$ , the process is indexable, i.e., for any belief  $x$ , if  $V_m(x; a=0) \geq V_m(x; a=1)$ , then  $V_m(x; a=0) \geq V_m(x; a=1), \forall m \geq m$ .*

#### 4.1.6 Numerical evaluation of indexability

**Proposition 1.** *If  $m < R(0)\beta \frac{R(n_o-1)-R(0)}{1-\beta}$ ,  $\Phi(m) = \emptyset$ ; if  $m > R(n_o-1)$ ,  $\Phi(m)$  is the whole belief state space.*

Thus, it should be determined whether the set  $\Phi(m)$  monotonically increases for  $m \subseteq \left[ R(0)\beta \frac{R(n_o-1)-R(0)}{1-\beta}, R(n_o-1) \right]$ . Numerically, this limited  $m$  range is discretized and then it is evaluated if  $\Phi(m)$  monotonically increases with the increase of discretized  $m$ . Given the subsidy  $m$ ,  $\Phi(m)$  can be determined by solving a special POMDP model whose conditional observation probability is dependent on start state and action. The algorithm returns a set  $D$  which contains  $n_s$ -length vectors  $d_1, d_2, \dots, d_{|D|}$ . Every vector  $d_i$  is associated with an optimal action  $e_i$ . Given the belief  $b$ , the optimal action is determined by  $a^{opt} = e_i, i = \arg \max_j b^T d_j$ . Thus,  $\Phi(m) = \cup_{i:e_i=0} \{b : b^T d_i \geq b^T d_j, \forall j\}$ .

Given  $m_0 < m_1$ , our aim is to check whether  $\Phi(m_0) \subseteq \Phi(m_1)$ . A MILP to verify whether such condition holds can be solved.

#### 4.1.7 Computation of Whittle index Policy

Given the indexability, Whittle index can be found by doing a binary search within the range  $m \subseteq \left[ R(0)\beta \frac{R(n_o-1)-R(0)}{1-\beta}, R(n_o-1) \right]$ . Given the upper bound  $ub$  and lower bound  $lb$ , the problem with middle point  $\frac{lb+ub}{2}$  as passive subsidy is sent to the special POMDP solver to find the optimal action for the current belief. If the optimal action is active, then the Whittle index is greater than the middle point so  $lb \leftarrow \frac{lb+ub}{2}$  or else  $ub \leftarrow \frac{lb+ub}{2}$ . This binary search algorithm can find Whittle index with arbitrary precision. Naively, the Whittle index policy can be found by computing the  $\epsilon$ -precision indices of all arms and then picking the  $k$  arms with the highest indices.

Specifically, let  $A$  be the Whittle index policy to be returned: it is set to be  $\emptyset$  at the beginning.  $S$  is the set of arms that are not known whether belonging to  $A$  or not and is set to be the whole set of arms at the beginning. Before it finds top- $k$  arms, it tests all the arms in  $S$  about their optimal action with subsidy  $\frac{lb+ub}{2}$ . If the optimal action is 1, it means this arm's index is higher

than  $\frac{lb+ub}{2}$  and it is added to  $S_1$ ; if the optimal action is 0, it means this arm's index is lower than  $\frac{lb+ub}{2}$  and it is added to  $S_0$ . At this moment, all arms in  $S_1$  have higher indices than all arms in  $S_0$ . If there is enough space in  $A$  to include all arms in  $S_1$ ,  $S_1$  is added to  $A$ , remove them from  $S$  and set the upper bound to be  $\frac{lb+ub}{2}$  because  $S_1$  belongs to Whittle index policy set and all the rest arms have the index lower than  $\frac{lb+ub}{2}$ . If there is not enough space in  $A$ ,  $S_0$  is removed from  $S$  and set the lower bound to be  $\frac{lb+ub}{2}$  because  $S_0$  does not belong to Whittle index policy set and all the rest arms have the index higher than  $\frac{lb+ub}{2}$ .

#### 4.1.8 Special POMDP formulation

Here, the algorithm to compute the passive action set  $\Phi(m)$  with the subsidy  $m$  is discussed. This problem can be viewed as solving a special POMDP model whose conditional observation probability is dependent on start state and action while the conditional observation probability is dependent on end state and action in standard POMDPs.

The original state is  $s$ , the agent takes action  $a$ , and the state transitions to  $s'$  according to  $P(s'|s, a)$ . However, the observation  $o$  the agent gets during this process is dependent on  $s$  and  $a$  in our special POMDPs; while it depends on  $s'$  and  $a$  in standard POMDPs.

The special POMDP formulation for our problem is straightforward.

- The state space is  $\mathcal{S} = \{0, 1, \dots, n_s - 1\}$ .
- The action space is  $\mathcal{A} = \{0, 1\}$ , where  $a = 0$  represents passive action (do not protect) and  $a = 1$  represents active action (protect).
- The observation space is  $\mathcal{O} = \{-1, 0, 1, \dots, n_o - 1\}$ . It adds a *fake* observation  $o = -1$  to represent no observation when taking action  $a = 0$ . It's called fake because defenders have probability 1 to observe  $o = 1$  no matter what the state is when they take action  $a = 0$ , so this observation does not provide any information. When defenders take action  $a = 1$ , they may observe observations  $\mathcal{O} \setminus \{-1\}$ .
- The conditional transition probability  $P(s'|s, a)$  is defined to be  $P(s' = j|s = i, a = 1) = T_{ij}^1$  and  $P(s' = j|s = i, a = 0) = T_{ij}^0$ .
- The conditional observation probability  $P(o|s, a)$  is defined to be  $P(o = -1|s, a = 0) = 1, \forall s \in \mathcal{S}; P(o = j|s = i, a = 1) = \mathcal{O}_{ij}$ . Note that the conditional observation probability here is dependent on the start state  $s$  and action  $a$ , while it depends on end state  $s$  and action  $a$  in standard POMDP models. Intuitively, defenders' observation of attack activities today depends on the attack intensity today, not the transitioned attack intensity tomorrow.

- The reward function  $R$  is:

$$R(s, s', a, o) = \begin{cases} 0, & a = 0 \\ R(o), & a = 1 \end{cases}$$

With the transition probability and observation probability,  $R(s, a)$  can be computed. Note that this formulation is also slightly different due to the different definition of observation probability.

$$R(s, a) = \sum_{s' \in \mathcal{S}} P(s'|s, a) \sum_{o \in \mathcal{O}} P(o|s, a) R(s, s', a, o)$$

#### 4.1.9 Value iteration for the special POMDP

Different from standard POMDP formulation, the belief update in the special POMDP formulation is:

$$b'(s') = \frac{\sum_{s \in \mathcal{S}} b(s) P(o|s, a) P(s'|s, a)}{P(o|b, a)}$$

where

$$P(o|b, a) = \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} b(s) P(o|s, a) P(s'|s, a) = \sum_{s \in \mathcal{S}} b(s) P(o|s, a)$$

Similar to standard POMDP formulation, the value function is:

$$V'(b) = \max_{a \in \mathcal{A}} \left( \sum_{s \in \mathcal{S}} b(s) R(s, a) + \beta \sum_{o \in \mathcal{O}} P(o|b, a) V(b_a^o) \right)$$

which can be broken up to simpler combinations of other value functions:

$$\begin{aligned} V'(b) &= \max_{a \in \mathcal{A}} V_a(b) \\ V_a(b) &= \sum_{o \in \mathcal{O}} V_a^o(b) \\ V_a^o(b) &= \frac{\sum_{s \in \mathcal{S}} b(s) R(s, a)}{|\mathcal{O}|} + \beta P(o|b, a) V(b_a^o) \end{aligned}$$

All the value functions can be represented as  $V(b) = \max_{\alpha \in \mathcal{D}} b \alpha$  since the update process maintains this property, so the set  $\mathcal{D}$  is updated when updating the value function.

#### 4.1.10 Planning from POMDP view

Every single target can be modeled as a special POMDP model. Given that, these POMDP models at all targets can be combined to form a special POMDP model that describe the whole problem, and solving this special POMDP model leads to defenders' exact optimal strategy. Use the superscript  $i$  to denote target  $i$ . Generally, the POMDP model for the whole problem is the cross product of the single-target POMDP models at all targets with the constraint that only  $k$  targets are protected at every round.

- The state space is  $\mathbf{S} = \mathbf{S}^1 \times \mathbf{S}^2 \times \dots \times \mathbf{S}^n$ . Denote  $s = (s^1, s^2, \dots, s^n)$ .
- The action space is  $\mathbf{A} = \{(a^1, a^2, \dots, a^n) | a^j \in \{0, 1\}, \forall j \in \mathbb{N}, \sum_{j \in \mathbb{N}} a^j = k\}$ , which represents that only  $k$  targets can be protected at a round. Denote  $a = (a^1, a^2, \dots, a^n)$ .
- The observation space  $\mathbf{O} = \mathbf{O}^1 \times \mathbf{O}^2 \times \dots \times \mathbf{O}^n$ . Denote  $o = (o^1, o^2, \dots, o^n)$ .
- The conditional transition probability is  $P(s'|s, a) = \prod_{j \in \mathbb{N}} P^j(s'^j | s^j, a^j)$ .
- The conditional observation probability is  $P(o|s, a) = \prod_{j \in \mathbb{N}} P^j(o^j | s^j, a^j)$ .
- The reward function is  $R(s, s', a, o) = \prod_{j \in \mathbb{N}} R(s^j, s'^j, a^j, o^j)$ .

Silver and Veness [20] have proposed POMCP algorithm, which provides high quality solutions and is scalable to large POMDPs. The POMCP algorithm only requires a simulator of the problem so it also applies to our special POMDPs. At a high level, the POMCP algorithm is composed of two parts: it uses a particle filter to maintain an approximation of the belief state; it draw state samples from the particle filter and then use MCTS to simulate what will happen next to find the best action. It uses a particle filter to approximate the belief state because it is even computationally impossible in many problems to update belief state due to the extreme large size of the state space. However, in our problem, the all-target POMDP model is the cross product of the single-target POMDP models at all targets. The single-state POMDP model is small so that it is computationally inexpensive to maintain its belief state. Thus, the state  $s^i$  at target  $i$  can be sampled from its belief state and then compose them together to get the state sample  $s = (s^1, s^2, \dots, s^n)$  for the all-target POMDP model.

## 4.2 Online planning for optimal protector strategies in resource conservation games

Protectors (law enforcement agencies) try to protect natural resources, while extractors (criminals) seek to exploit them. In many domains, such as illegal fishing, the extractors know more about the distribution and richness of the resources than the protectors, making it extremely difficult for the protectors to optimally allocate their assets for patrol and interdiction. Fortunately, extractors carry out frequent illegal extractions, so protectors can learn about the richness of resources by observing the extractor's behavior.



In resource conservation domains, the protector often does not know the distribution of resources while the extractor may have more information about it, e.g. preventing illegal fishing. Our goal is to provide an optimal asset deployment (e.g., patrol) strategy for the protector, given her lack of knowledge about the distribution of resources.

#### 4.2.1 Problem formulation

In resource conservation games, the extractor’s frequent illegal extractions provide the protector with the opportunity to learn about the distribution of resources by observing the extractor’s behavior. The aim is the construction of an online policy for the protector to maximize her utility given observations of the extractor. At every round, the protector chooses one site to protect and the extractor simultaneously chooses one site to steal from. Both the extractor and the protector have full knowledge about each other’s previous actions.

In our model, the amount of resources at each site will be fixed and the extractor will have full knowledge of this distribution. The protector will have to learn this distribution by observing the extractor’s behavior.

There is a finite time horizon  $t \in T$ . There are  $n$  sites indexed by  $N = \{1, 2, \dots, n\}$  that represent the locations of the natural resource in question: the extractor wants to steal resources from these sites and the protector wants to interdict the extractor. The value of the sites to the extractor is represented in terms of their utilities. Each site has a utility  $u(i)$  that is only known to the extractor. The utility space is discretized into  $m$  levels,  $u(i) \in M = \{1, 2, \dots, m\}$ . Human beings cannot distinguish between tiny differences in utilities in the real world, so discretizing these utilities is justified. For  $n$  sites and  $m$  utility levels, there are  $m^n$  possible sets of utilities across all sites. The distribution of resources is then captured by the vector of utilities at each site, and the set of possible resource distributions is:

$$U = \{(u(1), u(2), \dots, u(n)) : u(i) \in M, \forall i \in N\}$$

Assuming that the resource levels  $u(i)$ ,  $i \in N$  are independent from each other, at the beginning of the game, the protector may have some prior knowledge about the resource levels  $u(i)$  at each site  $i \in N$ . This prior knowledge is represented as a probability density function  $p(u(i))$  over  $M$ . If the protector does not know anything about  $u(i)$ , then a uniform prior for  $u(i)$  over  $M$  is adopted. At each time  $t \in T$ , the protector chooses a site  $a_t \in N$  to protect and the extractor simultaneously chooses a site  $o_t \in N$  from which to steal. If  $a_t = o_t$ , the protector catches the extractor and the extractor is penalized by the amount  $P(o_t) < 0$ ; if  $a_t \neq o_t$ , the extractor successfully steals resources from site  $o_t$  and gets a payoff of  $u(o_t)$ . For clarity, the protector’s interdiction is always successful whenever it visits the same site as the extractor. Additionally, the protector fully observes the moves of the extractor, likewise, the extractor fully observes the moves of the protector. Note that the penalty  $P(i)$ ,  $i \in N$  is known to both the protector and the extractor. A zero-sum game is adopted,

so the protector is trying to minimize the extractor's payoffs. In most resource conservation domains, the extractor pays the same penalty  $P$  if he is seized independent of the site he visits. Varying penalties across sites for greater generality are allowed. A fictitious Quantal Response playing (FQR) extractor is assumed. Specifically, a fictitious extractor assumes the protector's empirical distribution will be his mixed strategy in the next round. In this behavior model, the extractor makes decisions based on the parameters  $u(i), P(i), i \in N$ , as well as the protector's actions in previous rounds.

The extractor behaves in the following way: in every round, he computes the empirical coverage probability  $c_i$  for every site  $i$  based on the history of the protector's actions, then computes the expected utility  $EU(i) = c(i)P(i) + (1 - c(i))u(i)$  for every site and finally attempts to steal from the site  $i$  with the probability proportional to  $e^{\lambda EU(i)}$  where  $\lambda \geq 0$  is the parameter representing the rationality of the player (higher  $\lambda$  represents a more rational player).

To implement the above model, two technical questions must be resolved. First, at every round  $t$ , based on her current belief about  $u$ , how should the protector choose sites to protect in the next round? Second, after each round, how should the protector use the observation of the latest round to update her beliefs about  $u$ ? Here, decision making and belief updating in a partially observable environment, where the payoffs  $u$  are unobservable and the extractor's actions are observable, are being studied, which is the exact setup for a POMDP. A two-player game is now setup as a POMDP  $\{S, A, O, T, \Omega, R\}$  where the extractor follows a quantal response model.

- The state space of our POMDP is  $S = U \times Z^n$ , which is the cross product of the utility space and the count space.  $U$  is the utility space.  $Z^n$  is the set of possible counts of the protector's visits to each site, where  $C_t \in Z^n$  is an integer-valued vector where  $C_t(i), i \in N$  is the number of times that the protector has protected site  $i$  at the beginning of round  $t \in T$ . A particular state  $s \in S$  is written as  $s = (u, C)$ , where  $u$  is the vector of utility levels for each site and  $C$  is the current state count. The initial beliefs are expressed by a distribution over  $s = (u, 0)$ , induced by the prior distribution on  $u$ .  $c_t(i) = \frac{C_t(i)}{t-1}$  denotes the frequency with which the protector visits site  $i$  at the beginning of round  $t \in T$ .  $c_1 = 0$  by convention.
- The action space  $A$  is  $N$ , representing the site the protector chooses to protect.
- The observation space  $O$  is  $N$ , representing the site the extractor chooses to attempt to steal from.
- Let  $e_a \in R^n$  denote the unit vector with a 1 in slot  $a \in N$  and zeros elsewhere. The conditional transition probability  $T$  governing the evolution of the state is:

$$T(s' = (u', C') | s = (u, C), a) = \begin{cases} 1 & u = u', C' = C + e_a, \\ 0, & \text{otherwise} \end{cases}$$

Specifically, the evolution of the state is deterministic. The underlying utilities do not change, and the count for the site visited by the protector increases by one while all others stay the same.

- $EU(u, C) \in R^n$  defines the vector of empirical expected utilities for the extractor for all sites when the actual utility is  $u$  and the count is  $C$ ,

$$[EU(u, C)](i) = c(i)P(i) + (1 - c(i))u(i), \forall i \in N,$$

when  $t \geq 1$ .  $[EU(u, 0)](i) = u(i)$  by convention. Hence, our observation probabilities are explicitly:

$$\Omega(o | s' = (u, C), a) = \frac{e^{\lambda[EU(u, C - e_a)](o)}}{\sum_{i \in N} e^{\lambda[EU(u, C - e_a)](i)}}$$

the probability of observing the extractor takes action  $o$  when the protector takes action  $a$  and arrives at state  $s$ . Note that both  $a$  and  $o$  are the actions the protector/extractor take at the same round.

- The reward function  $R$  is:

$$R(s = (u, C), s' = (u, C + e_a), a, o) = \begin{cases} -P(o), & a = o \\ -u(o), & a \neq o \end{cases}$$

#### 4.2.2 GMOP algorithm

The size of the utility space  $U$  is  $m^n$ , and the size of the count space is  $O\left(\frac{T^n}{n!}\right)$ . The computational cost of the latest POMDP solvers soon become unaffordable for us as the problem size grows.

Silver and Veness [20] have proposed the POMCP algorithm, which provides high quality solutions for large POMDPs. The POMCP algorithm uses a particle filter to approximate the belief state. Then, it uses Monte Carlo tree search (MCTS) for online planning where state samples are drawn from the particle filter and the action with the highest expected utility based on Monte Carlo simulations is chosen. However, the particle filter is only an approximation of the true belief state and is likely to move further away from the actual belief state as the game goes on, especially when most particles get depleted and new particles need to be added. Adding new particles will either make the particle filter a worse approximation of the exact belief state, if the added particles do not follow the distribution of the belief state or be as difficult as drawing samples directly from the belief state, if the added particles do follow the distribution of the belief state.

Our POMDP has specific structure that can be exploited. The count state in  $S$  is known and the utility state does not change, making it possible to draw samples directly from the exact belief state using Gibbs sampling. We propose the GMOP algorithm that draws samples directly from the exact belief state using Gibbs sampling, and then runs MCTS. The samples drawn directly from the belief state better represent the true belief state compared to samples drawn from a particle filter.

At a high level, in round  $t$  the protector draws samples of state  $s$  from its belief state  $B_t(s)$  using Gibbs sampling and then it runs MCTS using those samples. Finally, it executes the action with the highest expected utility. MCTS starts with a tree that only contains a root node. Since the count state  $C_t$  is already known, the protector only needs to sample the utility state  $u$  from  $B_t$ . The sampled state  $s$  is comprised of the sampled utility  $u$  and the count  $C_t$ .

Gibbs sampling [2] is a Markov chain Monte Carlo (MCMC) algorithm for sampling from multivariate probability distributions. Let  $X = (x_1, x_2, \dots, x_n)$  be a general random vector with  $n$  components and with finite support described by the multivariate probability density  $p(X)$ . Gibbs sampling only requires the conditional probabilities  $p(x_i|x_{-i})$  to simulate  $X$ , where  $x_{-i} = (x_j)_{j \neq i}$  denotes the subset of all components of  $X$  except component  $i$ . Gibbs sampling is useful when direct sampling from  $p(X)$  is difficult. Suppose  $k$  samples of  $X = (x_1, x_2, \dots, x_n)$  should be obtained.

Gibbs sampling works in general to produce these samples using only the conditional probabilities  $p(x_i|x_{-i})$ . It constructs a Markov chain whose steady-state distribution is given by  $p(X)$ , so that the samples also follow the distribution  $p(X)$ . The states of this Markov chain are the possible realizations of  $X = (x_1, x_2, \dots, x_n)$ , and a specific state  $X_i$  is denoted as  $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$  (there are finitely many such states by our assumption). The transition probabilities of this Markov chain,  $Pr(X_j|X_i)$ , follow from the conditional probabilities  $p(x_i|x_{-i})$ . Specifically,  $Pr(X_j|X_i) = p(x_l|x_{-l})$  when  $x_{jv} = x_{iv}$  for all  $v$  not equal to  $l$ , and is equal to zero otherwise, i.e. the state transitions only change one component of the vector-valued sample at a time. This Markov chain is reversible (meaning  $p(X_i)Pr(X_j|X_i) = p(X_j)Pr(X_i|X_j), \forall i, j$ ) so  $p(X)$  is its steady-state distribution.

### 4.2.3 Applying Gibbs sampling in GMOP

Let  $B_t$  be the probability distribution representing the protector's beliefs about the true utilities at the beginning of round  $t \geq 1$ ;  $B_1$  represents the protector's prior beliefs when the game starts. Let  $B_t(u)$  denote the probability of the vector of utilities  $u$  with respect to the distribution  $B_t$ . Let  $B$  be the prior belief distribution and  $B$  be the posterior belief distribution. Our Bayesian belief update rule to obtain  $B$  from  $B$  and the observation is explicitly:

$$\begin{aligned}
B'(s' = (u, C)) &= \eta \Omega(o|s', a) \sum_{s \in S} T(s'|s, a) B(s) \\
&= \eta \Omega(o|s', a) B(s = (u, C - e_a))
\end{aligned}$$

If  $a_t$  and  $o_t$  represent the actions that the protector and the extractor choose to take at round  $t$ :

$$\begin{aligned}
B_t(u) &= \eta B_{t-1}(u) \Omega(o_{t-1}|s = (u, C_t), a_{t-1}) \\
&= \eta' B_1(u) \prod_{i=1}^{t-1} \Omega(o_i|s = (u, C_{i+1}), a_i)
\end{aligned}$$

It follows that the posterior belief  $B_t$  is proportional to the prior belief  $B_1$  multiplied by the observation probabilities over the entire history. Since there are  $mn$  possible utilities, it is impossible to store and update  $B_t$  when  $m$  and  $n$  are large, and thus it is impossible to sample directly from  $B_t$ . Hence, Gibbs sampling is adopted.

Only the conditional probabilities  $p(u_i|u_{-i})$ ,  $\forall i$  in  $B_t$  are needed:

$$\begin{aligned}
p(u_i|u_{-i}) &= \eta p(u_i, u_{-i}) = \eta B_t(u_i, u_{-i}) = \\
&\eta' B_1(u_i, u_{-i}) \prod_{j=1}^{t-1} \Omega(o_j|s = (u = (u_i, u_{-i}), C_{j+1}), a_j) = \\
&\eta'' B_1(u_i) \prod_{j=1}^{t-1} \Omega(o_j|s = (u = (u_i, u_{-i}), C_{j+1}), a_j)
\end{aligned}$$

This quantity is easy to compute where  $B_1(u_i)$  is the prior probability that site  $i$  has utility  $u_i$ . Besides the conditional probability, also a valid  $u$  with  $B_t(u) > 0$  should be found to initialize Gibbs sampling. Finding such a  $u$  is easy in our FQR model because any  $u$  with  $B_1(u) > 0$  satisfies  $B_t(u) > 0$ .

## References

- [1] Matthew Brown, William B Haskell, and Milind Tambe. Addressing scalability and robustness in security games with multiple boundedly rational adversaries. In *Decision and Game Theory for Security*, pages 23–42. Springer, 2014.
- [2] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.
- [3] Fei Fang, Peter Stone, and Milind Tambe. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.
- [4] Nika Haghtalab, Fei Fang, Thanh H Nguyen, Arunesh Sinha, Ariel D Procaccia, and Milind Tambe. Three strategies to success: Learning adversary models in security games.
- [5] William B Haskell, Debarun Kar, Fei Fang, Milind Tambe, Sam Cheung, and Elizabeth Denicola. Robust protection of fisheries with compass. In *AAAI*, pages 2978–2983, 2014.
- [6] Debarun Kar, Fei Fang, Francesco Delle Fave, Nicole Sintov, and Milind Tambe. A game of thrones: when human behavior models compete in repeated stackelberg security games. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1381–1390. International Foundation for Autonomous Agents and Multiagent Systems, 2015.
- [7] Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.
- [8] Thanh H Nguyen, Francesco M Delle Fave, Debarun Kar, Aravind S Lakshminarayanan, Amulya Yadav, Milind Tambe, Noa Agmon, Andrew J Plumptre, Margaret Driciru, Fred Wanyama, et al. Making the most of our regrets: Regret-based solutions to handle payoff uncertainty and elicitation in green security games. In *Decision and Game Theory for Security*, pages 170–191. Springer, 2015.
- [9] Thanh H Nguyen, Arunesh Sinha, Shahrzad Gholami, Andrew Plumptre, Lucas Joppa, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Rob Critchlow, et al. Capture: A new predictive anti-poaching tool for wildlife protection. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 767–775. International Foundation for Autonomous Agents and Multiagent Systems, 2016.
- [10] Thanh H Nguyen, Arunesh Sinha, and Milind Tambe. Addressing behavioral uncertainty in security games: An efficient robust strategic solution for defender patrols.

- [11] Thanh H Nguyen, Arunesh Sinha, and Milind Tambe. Conquering adversary behavioral uncertainty in security games: An efficient modeling robust based algorithm. 2016.
- [12] Thanh Hong Nguyen, Rong Yang, Amos Azaria, Sarit Kraus, and Milind Tambe. Analyzing the effectiveness of adversary modeling in security games. In *AAAI*, 2013.
- [13] Christos H Papadimitriou and John N Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.
- [14] James Pita, Manish Jain, Janusz Marecki, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Deployed armor protection: the application of a game theoretic model for security at the los angeles international airport. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: industrial track*, pages 125–132. International Foundation for Autonomous Agents and Multiagent Systems, 2008.
- [15] Yundi Qian, William B Haskell, Albert Xin Jiang, and Milind Tambe. Online learning and planning in resource conservation games.
- [16] Yundi Qian, William B Haskell, Albert Xin Jiang, and Milind Tambe. Online planning for optimal protector strategies in resource conservation games. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 733–740. International Foundation for Autonomous Agents and Multiagent Systems, 2014.
- [17] Yundi Qian, Chao Zhang, Bhaskar Krishnamachari, and Milind Tambe. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 123–131. International Foundation for Autonomous Agents and Multiagent Systems, 2016.
- [18] Global Tiger Initiative Secretariat. Global tiger recovery program implementation plan: 2013-14. *Report, The World Bank, Washington, DC*, 2013.
- [19] Eric Shieh, Bo An, Rong Yang, Milind Tambe, Craig Baldwin, Joseph DiRenzo, Ben Maule, and Garrett Meyer. Protect: A deployed game theoretic system to protect the ports of the united states. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 13–20. International Foundation for Autonomous Agents and Multiagent Systems, 2012.
- [20] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In *Advances in neural information processing systems*, pages 2164–2172, 2010.

- [21] Marwan Simaan and Jose B Cruz Jr. On the stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications*, 11(5):533–555, 1973.
- [22] Arunesh Sinha, Debarun Kar, and Milind Tambe. Learning adversary behavior in security games: A pac model perspective. *arXiv preprint arXiv:1511.00043*, 2015.
- [23] Heinrich F. von Stackelberg. *Marktform und Gleichgewicht*, volume 1. Springer, 1934.
- [24] Peter Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pages 287–298, 1988.
- [25] Rong Yang, Benjamin Ford, Milind Tambe, and Andrew Lemieux. Adaptive resource allocation for wildlife protection against illegal poachers. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 453–460. International Foundation for Autonomous Agents and Multiagent Systems, 2014.
- [26] Rong Yang, Fernando Ordonez, and Milind Tambe. Computing optimal strategy against quantal response in security games. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 847–854. International Foundation for Autonomous Agents and Multiagent Systems, 2012.
- [27] Zhengyu Yin, Albert Xin Jiang, Milind Tambe, Christopher Kiekintveld, Kevin Leyton-Brown, Tuomas Sandholm, and John P Sullivan. Trusts: Scheduling randomized patrols for fare inspection in transit systems using game theory. *AI Magazine*, 33(4):59, 2012.